Introduction to Data Science

BRIAN D'ALESSANDRO

ADJUNCT PROFESSOR, NYU

FALL 2018

Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute, except for as needed as a pedagogical tool in the subject of Data Science.

THIS COURSE'S #1 GOAL IS TO HELP YOU TO

BECOME A DATA SCIENTIST

2018 – STILL SEXY?

THE MAGAZINE

October 2012

Harvard Business Review

ARTICLE PREVIEW To read the full article, **sign-in** or **register**. HBR subscribers, click **here to register for FREE access »**

Data Scientist: The Sexiest Job of the 21st Century

*"Data scientists are the key to realizing the opportunities presented by big data. **They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions**.*

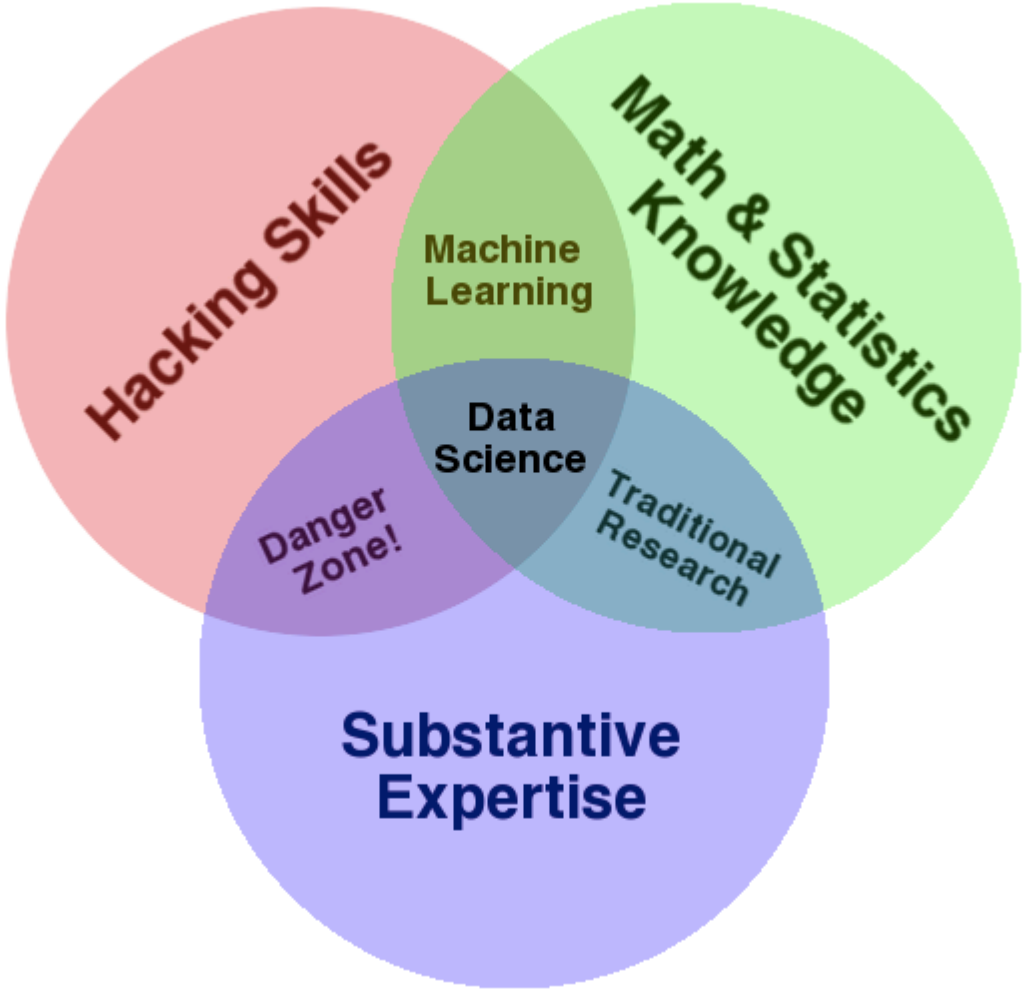
They find the story buried in the data and communicate it. And they don't just deliver reports:

They get at the questions at the heart of problems and devise creative approaches to them."

http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

THIS IS A POPULAR DIAGRAM

What skills do we expect in our data scientists?

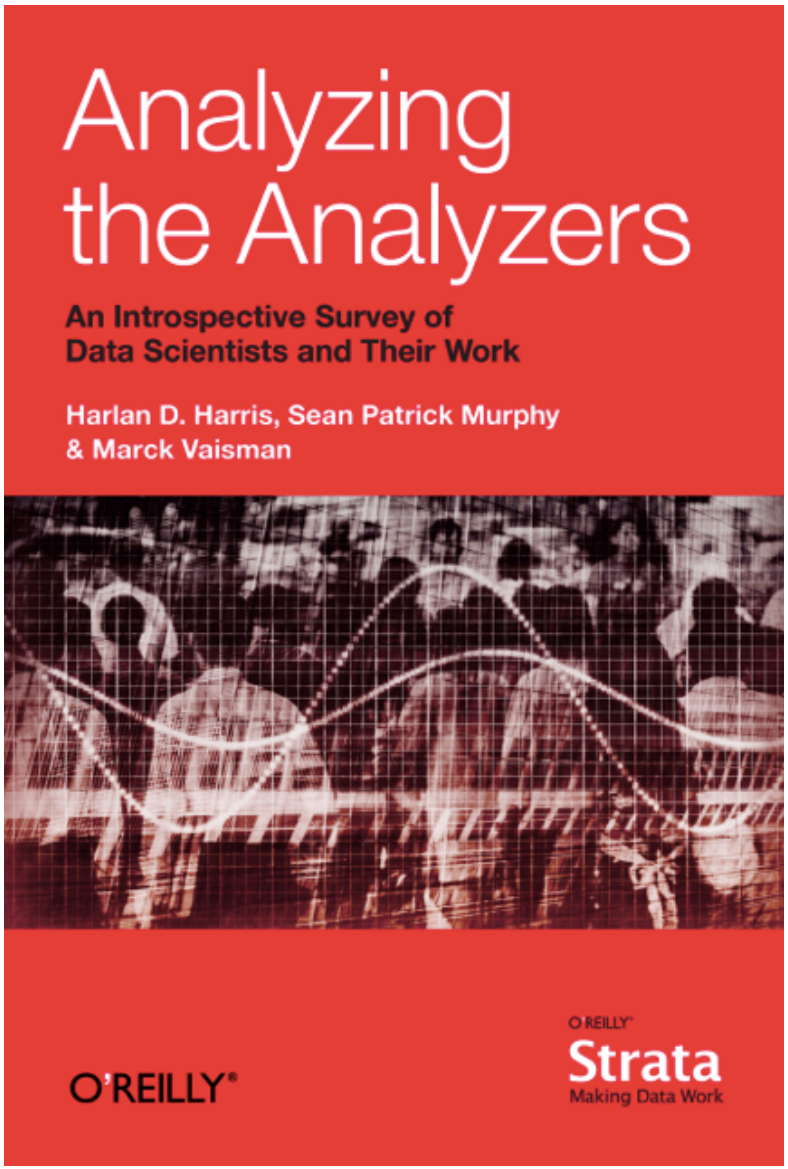


Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

TOWARDS A DEFINITION

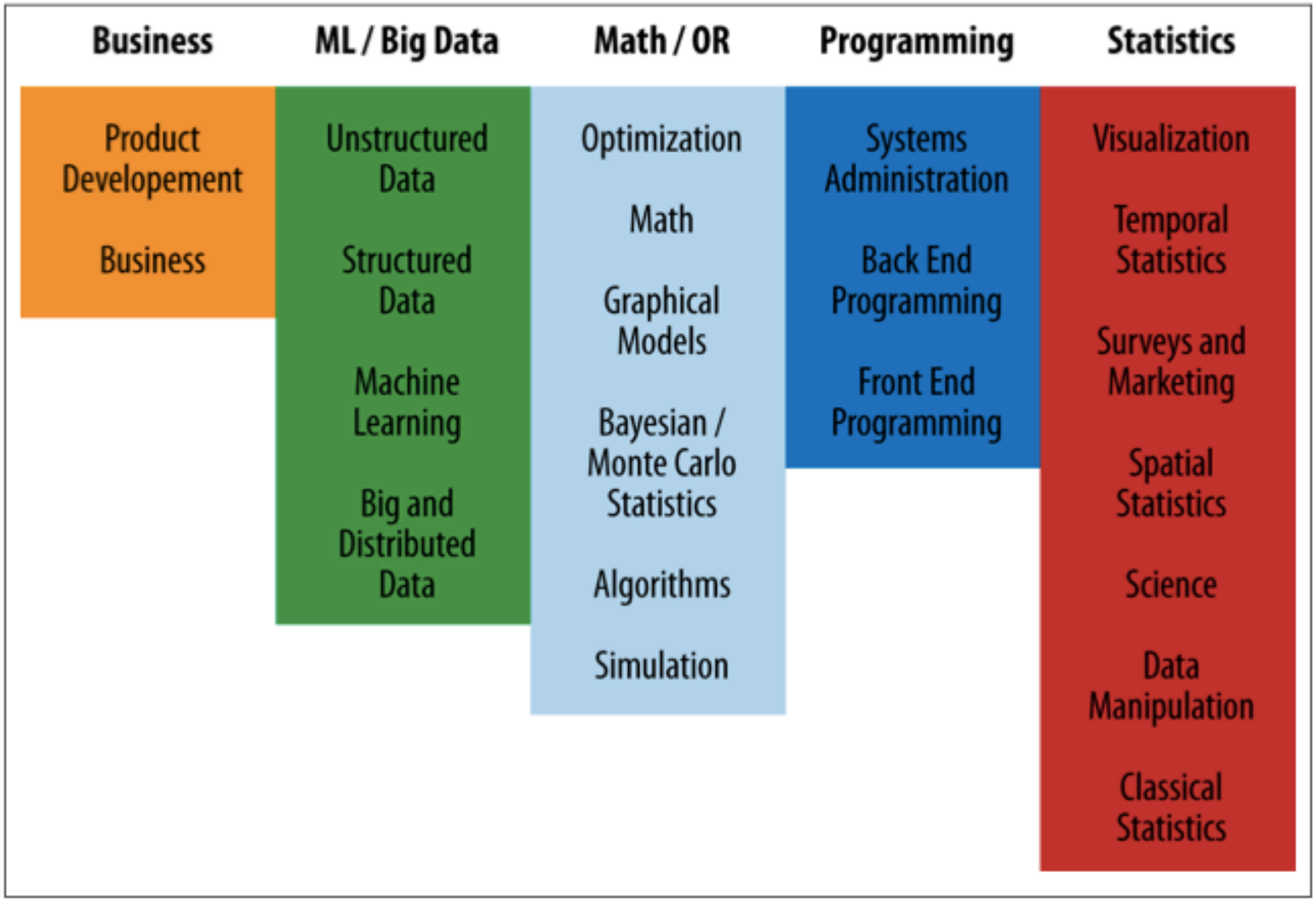
There is no ‘one-size-fits-all’ type of data scientist.

Luckily, people are using data science to define data science.



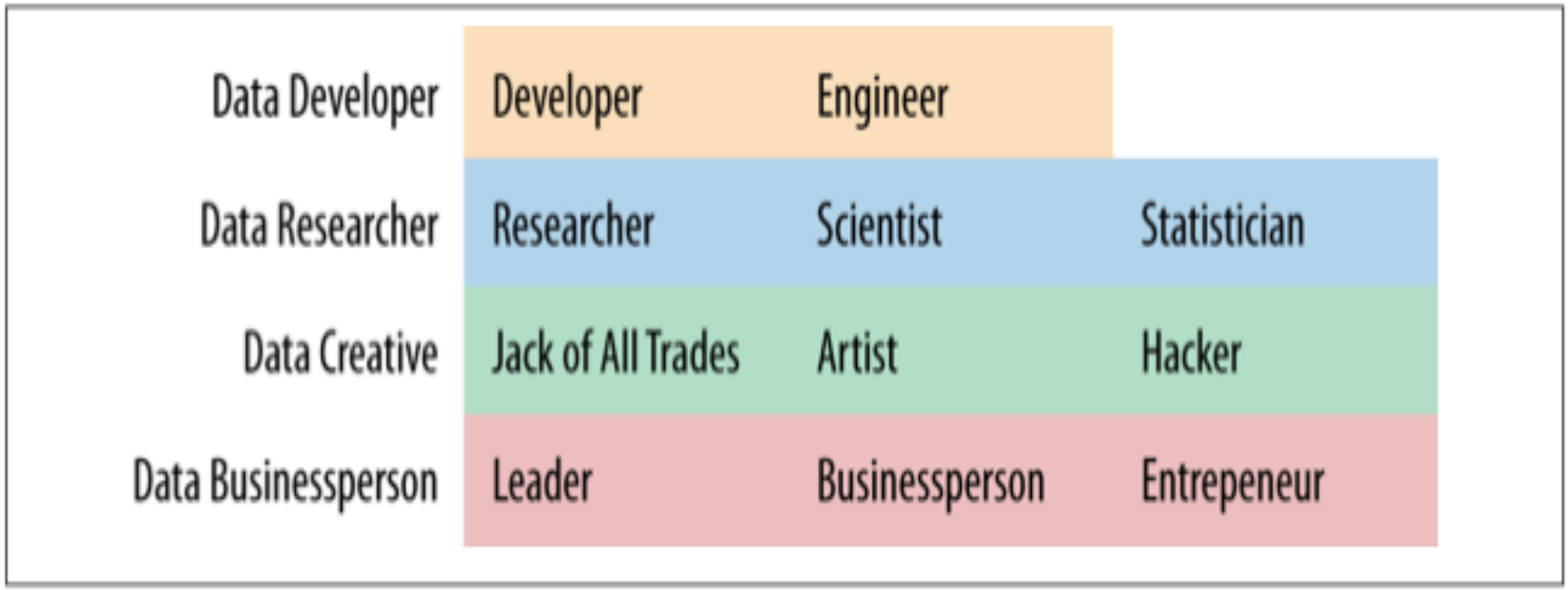
RANGE OF DS SKILLS

They're all very similar, but some categorization still helps.



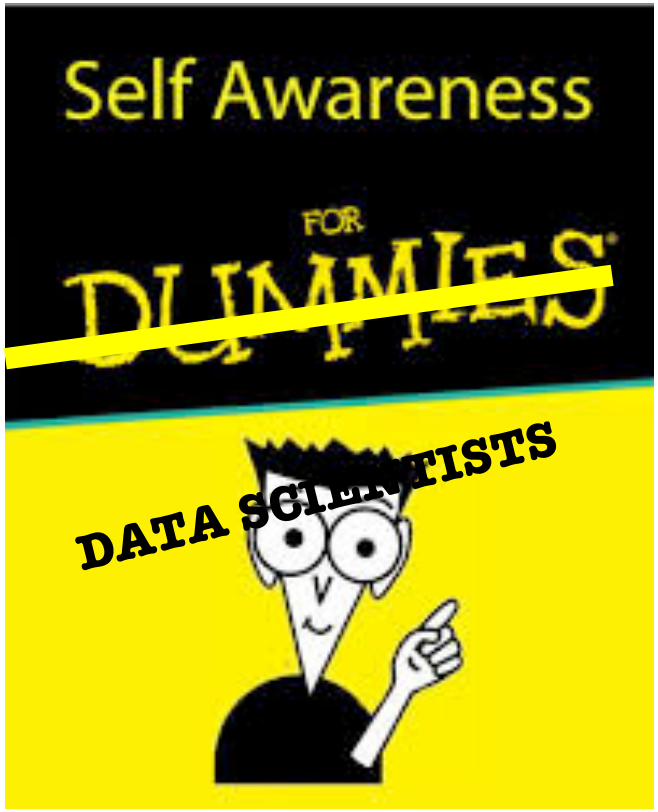
DATA ROLES

In Analyzing the Analyzers, the authors identified 4 types of “data scientists.”



IT MATTERS

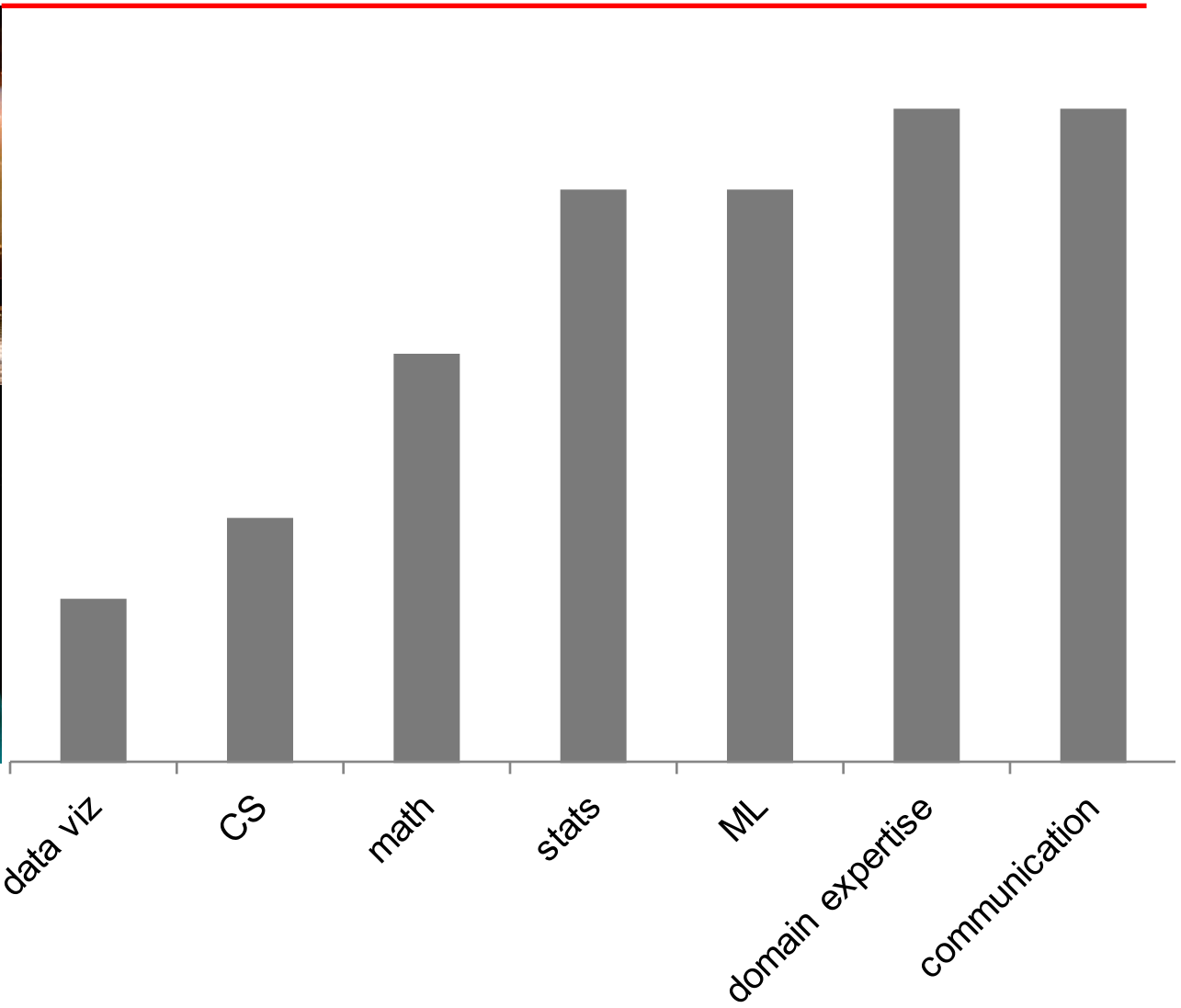
You don't have to fit into one bucket, but you should know where you are…



- Personal skills development
- Choosing the right job (your future boss might not know what a data scientist is, or should be)

DATA SCIENCE PROFILE

What I think I am…



WHY SCIENCE?

We defined 4 data roles, but what is the “science” of data science?

<u>The scientific method:</u> evaluating the merit of a hypothesis with rigorous empirical testing.

I.e.,

Given raw data, constraints and a problem statement, you have an infinite set of models to choose from, with which you will use to maximize performance on some evaluation metric, that you will have to specify.

Every design choice you make can be formulated as a hypothesis, upon which you will use rigorous testing and experimentation to either validate or refute.

BUT ITS STILL AN ART

Outside of modeling competitions, seldom is a well-posed problem and clean dataset presented to you.

Putting the art into your practice means…

- Translating problems into the language of data science
- Formulating reasonable hypotheses
- Developing an intuition for good vs. bad data, good vs. bad models.
- Abstracting problems to identify similarities
- Managing the DS process from end to end

REMINDER

With this course we want to emphasize the *soft* skills of data science

Art => Abstract and intuitive thinking

Science => Process

We'll cover necessary DS tools, but with the goal of applying them towards analytic problem solving.

THE RESUME: SKILLS

Through the lens of a Data Scientist Job Description

PhD is a proxy for:
- **experience**
- **research ability**
- **technical expertise**

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a
- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- Ability to communicate complex quantitative analysis and analytic approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

THE RESUME: SKILLS

Through the lens of a Data Scientist Job Description

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a rel
- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- Ability to communicate complex quantitative analysis and analytic approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

You can't be a Data Scientist if you can't handle data…

THE RESUME: SKILLS

Through the lens of a Data Scientist Job Description

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a relevant role
- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- Ability to communicate complex quantitative analysis and analytic approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

This is essentially the goal of this course.

THE HARD SKILLS

Through the lens of a Data Scientist Job Description

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a relevant role
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

Necessary: a scripting language, SQL and a scientific computing language. You will get hands-on experience with some of this in this course, and you should definitely develop these skills throughout this program.